文章编号：1003-0077（2017）00-0000-00

基于成分共享的英汉小句对齐语料库标注体系研究

葛诗利[1] 宋柔[1,2]

（1.广东外语外贸大学 外语研究与语言服务协同创新中心，广东省 广州市 510420；2.北京语言大学 信息科学学院，北京市 100083）

摘要：英汉小句对齐语料库服务于英语和汉语小句的语法结构对应关系研究和应用，对于语言理论和语言翻译（包括人的翻译和机器翻译）有重要意义。前人的语法理论和相关语料库的工作对于小句复合体和小句的界定缺乏充分研究，在理论上有缺陷，难以支持自然语言处理的应用。本文首先为英汉小句对齐语料库的建设做理论准备。从近年提出的汉语小句复合体的理论出发，本文界定了成分共享的概念，基于话头共享和引语共享来界定英语的小句和小句复合体，使小句和小句复合体具有功能的完整性和单一性。在此基础上，本文设计了英汉小句对齐的标注体系，包括英语 NT 小句标注和汉语译文生成及组合。语料库的标注表明，在小句复合体层面上英汉翻译涉及到的结构变换，其部件可以限制为英语小句和话头、话体，无需涉及话头和话体内部的结构。基于这些工作的英汉小句对齐语料库为语言本体研究和英汉语言对比、英汉机器翻译等应用提供了结构化的标注样本。

关键词：成分共享；话头共享；小句；小句复合体；英汉机器翻译

中图分类号：TP391 **文献标识码：**A

English-Chinese Clause Alignment Corpus Tagging System Based on Component Sharing

GE Shili[1] and SONG Rou[1,2]

(1.Guangdong University of Foreign Studies, Guangzhou, Guangdong 510420, China ; 2. Beijing Language and Culture University, Beijing 100083, China)

Abstract: English-Chinese clause alignment corpus serves the study and application of grammatical structure correspondence between English and Chinese clauses. It is of great significance to linguistic theory and language translation (including human translation and machine translation). Previous work on grammar theory and corpus lacks sufficient research on definitions of clause and clause complex. It is theoretically defective and difficult to support the application of natural language processing. Firstly, this paper makes theoretical preparations for the construction of English-Chinese clause alignment corpus. Starting from the theory of Chinese clause complex put forward in recent years, this paper defines the concept of component sharing, and further defines English clause and clause complex based on topic sharing and quotation sharing, which endows clause and clause complex with integrity and unity. Based on the study, an English-Chinese clause alignment annotation system is designed, including English NT clause tagging and Chinese translation generation and combination. The corpus annotation shows that, at the clause complex level, the components involved by structural transformation in English-Chinese translation can be limited to English clauses and related naming and telling, without involving the internal structure of namings and tellings. Based on these works, the English-Chinese clause aligned corpus provides research samples for linguistic ontology research, English-Chinese language comparison and English-Chinese machine translation.

收稿日期：2019-08-12；**定稿日期**：2019-08-26

基金项目：国家自然科学基金面上项目“NT 小句复合体模型的理论和应用研究”（61672175）；国家语委重点项目“英汉机器翻译译文错误分析及面向篇章的机器翻译解决方案研究”（ZDI135-30）

Key words: Component Sharing; Naming Sharing; Clause; Clause Complex; English-Chinese Machine Translation

1 英汉小句对齐语料库建设

语篇中有小句复合体、小句、短语、词和语素等语法层级。研究不同语言在各个层级上的对应关系对于语言学的理论和应用都具有重要意义。过去已有大量工作研究英语和汉语在小句层面上短语之间的对应关系，这方面的工作极大推动了英汉机器翻译的发展。但是，英语和汉语在小句复合体层面上小句之间的对应关系还少有研究。这种对应关系由于涉及的上下文范围大，机器自动捕捉困难，更需要人来归纳特征和规律。这是我们建设英汉小句对齐语料库的驱动要素。

英汉小句对齐语料库的原文来自宾州树库所用的华尔街日报。该语料库的建设包括 3 部分工作：将英语语篇切分为小句复合体，把小句复合体切分为小句，标注汉语译文内容与英语小句的对应关系。为此，在理论方面，需要界定英语的小句复合体和小句；在方法方面，需要设计英语语篇中切分出小句复合体和小句的标注体系，以及汉语译文内容与英语小句对应关系的标注体系。这些是本文要介绍的内容。

2 相关的语法理论和研究工作

在语法理论方面，最有影响的是形式语法和功能语法。形式语法的理论并没有专门研究英语小句复合体和小句的界定，更没有给出在语篇中切分出小句复合体和小句的操作方法。

基于形式语法的宾州英语树库，完全以句号、问号、叹号作为分割标记，在语篇中切分句子（即小句复合体）。大多数情况下，这样做没问题，但是在涉及引述语和引语时，可能发生错误。具体情况是：如果语篇内容的排列是引述语后面接有引语，引语中不止一个句子，引述语和引语之间没有标点或用逗号分开，宾州树库的处理方式是把引述语和引语中的第 1 句合成一个句子，引语中的其他句子分别划开。例如：

例 1

A Lorillard spokeswoman said , “This is an old story . We 're talking about years ago before anyone heard of asbestos having any questionable properties . There is no asbestos in our products now . ”

该句在宾州树库中被切分成 3 个句子[1]：

A Lorillard spokeswoman said , “This is an old story .

We 're talking about years ago before anyone heard of asbestos having any questionable properties .

There is no asbestos in our products now . ”

这样的切分打乱了引述语和引语之间的关联，显然不合适。

宾州树库对于句子内部的嵌套小句的结构，虽然做了句法类型标注，但没有划分出小句。

对比形式语法，功能语法的特点是更关注话语，关注语篇中小句复合体和小句层面，并且以功能分析为中心而不是以结构形式为中心。功能语法从消息功能、交际功能和表征功能这 3 种视角分析小句的功能，从配列方式和逻辑语义关系两个维度来分析小句复合体内小句间的关系。采用这一套体系，功能语法分析了英语的种种语法现象[2]。

由于功能语法在句法之外更多地考察了语篇和功能，所以适合于支持面向话语的语言应用研究。不同语言的句法结构形式可能有巨大差异，但消息功能、交换功能、表征功能则是相通的，因此功能语法体系很适合用于跨语言的研究和应用，包括语言比较、语言翻译等。我们进行英汉小句对齐语料库的建设，吸取了功能语法的重要思想。

但是，功能语法是一种语言理论，缺少面向话语的操作设计，特别是在小句的界定上存在问题，给应用带来困难。

第一，功能语法所界定的小句很多情况下功能不完整。

在功能语法的小句切分中，引述语和引语被完全分开。例 1 中 A Lorillard spokeswoman said 会被单独切成一个小句，于是其中及物动词 said

没有宾语，这个小句在句法上和功能上都不完整。

例 2

It did not distinguish fundamentally between the private and the public sector, which were treated as parts of a single whole. [2]

按照功能语法的观点，这是一个主从关系的小句复合体，逗号分开了两个小句：

It did not distinguish fundamentally between the private and the public sector,

which were treated as parts of a single whole.

从表征的视角看，第 2 个小句过程的参与者是第 1 小句中的 the private and the public sector，但它并未出现在第 2 个小句中，因此第 2 个小句的表征功能是不完整的。定语从句中的关系代词指代的对象受到句法约束，它一定是先行的名词短语的替身，但功能语法在小句界定中不考虑这种联系。

第二，功能语法所界定的小句很多情况下功能过多。

在功能语法中，名词短语后部的说明成分，无论是小句还是短语，都看作嵌进名词短语的低级阶成分，即嵌入性小句或嵌入性短语，不分离出来看。如此，功能语法的小句会把这种嵌入性成分所带有的功能收进小句自身的功能中去，使得小句功能过于复杂。虽然嵌入性成分所表现的功能与小句主干成分所表现的功能不在一个层次上，在主干层次上可以忽略，但是，不同语言的语法结构和表达方式不一样，有些语言未必有显在的级阶的区别，英语话语中的低级阶成分在汉语的同样意义的话语中未必是低级阶成分。遇到这种情况，功能语法小句中的低级阶成分就需要打破小句的界限而提升出来。

例 3

People who have enjoyed good educational opportunities ought to show it in their conduct and language.[3]

例中的定语从句 who have enjoyed good educational opportunities 单独看具有小句的形式，但在更大的话语环境中它是 people 的后置说明成分。功能语法依据英语的句法结构，把它看作嵌入到名词短语内部的成分，不是独立的小句，从而整个小句复合体只包含一个小句。于是，这个小句有两个主位述位结构、两个语气结构、两个过程结构，无论从消息、交换还是表征的视角来看功能都过于复杂。

汉语语法学家王力把这个句子译成汉语，他推荐的译文是：

一个人享受过良好的教育机会，应该在行为和语言上表现出来。[3]

在汉语译文中，这是两个小句，共享主语“一个人”，其中第 1 个小句对应的就是英语中的嵌入性小句。也就是说，英语的小句和汉语的小句不能对应，英语翻译成汉语的过程中需要把嵌入小句提升出来。

例 4

If industries were allowed to shrivel and fail they would cease producing the goods which the country exchanged for food (which it had ceased to produce for itself when it took the industrial option) and for the industrial raw materials which it did not possess within its own borders (now much reduced by loss of empire).[2]

这个例子按照功能语法的分析只有两个小句，第一个小句是条件从句：

If industries were allowed to shrivel and fail

第 2 个小句的主干部分是：

they would cease producing the goods

其余部分都是嵌入到 the goods 为中心语的名词短语中的修饰成分，但是这个修饰成分比例 3 还要复杂得多，在汉语译文中多处需要提升。下文中会仔细分析这个例子。

除了上述理论体系以外，已经发表的一些小句对齐语料库的工作也存在类似的问题。如日英小句对齐语料库[4]、保加利亚语-英语小句和句子对齐语料库[5]、汉英小句对齐语料库[6]对于日语、保加利亚语和汉语小句的界定都存在功能不全或过多的问题。

理想的小句在功能上应当是完整的、单一的，这样才适合用作小句复合体分析的基本单位，可以进行小句间逻辑语义分析、指代分析，便于做信息提取等应用。不同语言都用这样的小句作为基本功能单位，便可以建立统一的平台，进行信息比较和转换。

语法单位的界定涉及到语法体系上下层次的方方面面，又要考虑不同语言的差别，因此有

些不理想之处在所难免。但是，就小句复合体和小句的界定操作而言，只要允许不同小句复合体之间、不同小句之间可以共享成分，则上述问题都可以解决，从而可以方便地支持英汉的小句对齐操作。而且，基于这种成分共享的观念，可以建立起一个新的形式体系，揭示出一些隐含在话语中的语言性质，更有重要的理论意义。

3 成分共享

成分共享的现象在英语语法中很少被提及，汉语中有一些研究，但概念并未准确界定。

我们用 3 个特征来界定成分共享的概念：

（1）被共享的成分在文本中的字面上只出现一次，但在功能上被使用不止一次；

（2）被共享的成分出现位置和使用位置之间的关系有语法模式可循；

（3）被共享的成分若在使用位置处被补上，则往往造成冗余甚至歧义。

成分共享不同于省略。省略的成分可以被补上去，但认定应补的成分，不是靠语法模式，而是靠会话场景和语言外的知识。省略的成分补上后并无话语的冗余或歧义。

成分共享也不同于共指。同一个概念用相同或不同的词语表达，字面上出现不止一次，属于共指，不是共享。

例如：

He left the room and went to the playground.

这里 left the room 和 went to the playground 共享 he。这是两个谓语共享主语，是一种语法现象。如果把后一个谓语补上主语 he，则发生冗余：

He left the room and he went to the playground.

4 话头、话体、NT 小句及其标注

宋柔提出了话头（naming）和话体（telling）的概念，基于这对概念界定了汉语的小句，又基于话头共享结构界定了汉语的小句复合体[7]。本文将这一思想应用于英语。

话头是话语的字面上的出发点，话体是对于话头的陈述[8]。这对概念非常相似于功能语法中的主位和述位，区别在于，主位述位关系遵循英语的句法约束，话头话体关系可以打破某些句法结构的约束，目的是保证话头话体所成小句在功能上的完整性和单一性。

英语中的话头话体关系，有以下几种情况：

（1）主语和定式动词谓语

（2）先行语和定语从句

（3）先行语和后置的非定式动词短语/具有陈述意义的介词短语/具有解释性的名词短语/形容词短语/副词短语

（4）话头是句首的连词/副词短语/修饰性的介词短语，话体又是话头话体关系的结构。

上述的（1）和（4）同功能语法的主位述位关系一致，（2）和（3）是对功能语法的扩充。

话头共享有两种类型：一类是几个话体共用一个话头，这个话头在字面上只出现一次，上述 4 种情况都会导致这类共享。另一类是在某个话头的话体中的一个成分用作另外一个或多个话体的话头，上述（2）（3）两种情况会导致这类共享。

上述情况（4）表明，话头话体结构可能递归。最外层的话头与它的一个话体所组成的结构，称之为 NT 小句（naming-teling clause）。NT 小句不一定是句法上合规的小句，但经过机械性的变换就可以变成合规的小句。（注：本文说的合规的小句，大体上就是英语教学语法所描述的小句，也包括这种小句首部加连词而成的状语从句。NT 小句与合规小句的对应关系的系统性介绍见另文。）

下面对于第 2 节的部分实例，基于话头共享的概念予以标注。

例 2’

例 2 的话头共享结构表示如图 1。

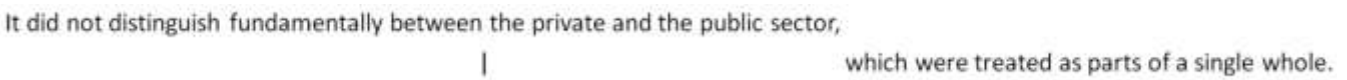

图 1 例 2 小句复合体的话头共享结构

上面第 1 行中主语 it 是话头，它的谓语 did not distinguish fundamentally between the private and the public sector 是话体。这个主谓结构合成一个 NT 小句，也是合规的小句。

第 2 行是一个定语从句，它的先行语是第 1 行的末尾的名词短语 the private and the public sector，它与这个定语从句是一对话头和话体，或者说定语从句共享第 1 个小句中的这个名词短语用作话头。这对话头话体组合成一个 NT 小句：

the private and the public sector + which were treated as parts of a single whole.

这个 NT 小句不是合规的小句，但只要经过一个机械性的变换，把关系代词删除，就是一个合规的小句：

the private and the public sector were treated as parts of a single whole.

在这个例句中，先行语与它的定语从句所成的话头话体关系不是主谓关系。我们在话头话体关系的标注中把这个定语从句换行列出，并把它缩进到它的话头即先行语的右端，话头的左边界则用竖线表示。我们称这种表示法为换行缩进图式。下面的例子都采用这种图式来表示话头话体关系。为减少标注工作，我们约定邻接的主谓关系所对应的话头话体关系不需要采用换行缩进方式来表示。

例 3’

例 3 的话头共享结构表示如下：

People
who have enjoyed good educational opportunities
ought to show it in their conduct and language.

图 2 例 3 小句复合体的话头共享结构

这个例子中，people 是话头，带有两个话体，一个是定语从句，另一个是定式动词谓语。它们形成 2 个 NT 小句：

People + who have enjoyed good educational opportunities

People + ought to show it in their conduct and language.

前一个去掉关系代词 who，就是合规的小句：

People have enjoyed good educational opportunities

第 2 个本身就是合规的小句。

例 4’

例 4 的话头共享结构表示如图 3。

If industries were allowed to shrivel and fail
they would cease producing the goods
| which the country exchanged for food
| (which it had ceased to produce for itself
when it took the industrial option)
and for the industrial raw materials
| which it did not process within its own borders
| (now much reduced by loss of empire).

图 3 例 4 小句复合体的话头共享结构

这个例子比较复杂，表示话头共享关系的换行缩进图式占据了 8 行。

首先是一个条件从句 If industries were allowed to shrivel and fail，它有两层话头话体关系。首层话头是连词 if，提示后面是条件。作为话体的条件中 industries 是过程参与者作话头，表示过程的 were allowed to shrivel and fail 是话体。这两层的话头话体关系构成正常合规的小句，在换行缩进图式中不换行。

例子的主句部分开始是一个主谓结构的小句 they would cease producing the goods，它的宾语的一部分 the goods 被后面的定语从句共享为话头。该定语从句是

which the country exchanged for food…and for the industrial raw materials…

两个介词短语的宾语 food 和 the industrial raw materials 又作为话头，带各自的定语从句作为话体。

food 的定语从句由主句和表示时间的状语从句构成，两个小句没有话头共享关系。

the industrial raw materials 的定语从句中，介

词宾语 its own borders 作为话头，带有动词的过去分词短语(now much reduced by loss of empire)做话体。

于是，这个例句有 8 个 NT 小句：

1) If industries were allowed to shrivel and fail

2) they would cease producing the goods

3) the goods +which the country exchanged for food

4) food + which it had ceased to produce for itself

5) when it took the industrial option

6) the goods +which the country exchanged + and for the industrial raw materials

7) the industrial raw materials + which it did not possess within its own borders

8) its own borders + now much reduced by loss of empire.

注意，第（6）个 NT 小句中第 2 个片段是未完的话体，接上第 3 个片段才是完整的话体。

这 8 个 NT 小句中，（1）（2）（5）本身就是合规的小句。（3）（4）（6）（7）的话头是逻辑宾语，这 4 个 NT 小句变成合规小句需要去掉关系代词和小句内部的连词，并且把话头调到宾语位置上去：

3) the country exchanged the goods for food

4) it had ceased to produce food for itself

6) the country exchanged the goods for the industrial raw materials

7) it did not possess the industrial raw materials within its own borders

第（8）个 NT 小句需加入 are 表示被动语态：

8）its own borders are now much reduced by loss of empire.

这 3 个例子显示，将话头和话体连接起来便得到 NT 小句，再经过简单机械的变换便得到合规的小句。NT 小句在结构和功能方面都既保证了完整性，又保证了单一性（即一个小句关于每一类功能只有一组结构），对于话语的理解、比较和翻译，是十分适当的单位。

我们定义的各种话体，其句法类型和功能的级阶不一样，在翻译中对于译文的影响也不一样。为此，在语料库的话头共享结构的换行缩进图式中，我们对于不同类型的话体标注了不同的类型符，从而可以研究话体的不同级阶对于译文的影响[9]。

5 基于引语共享的小句复合体标注

英语的小句复合体在书面上大部分情况下就是由句号、叹号、问号划分的。如本文第 2 节所述，这一方法在处理引述语和引语的关系时会造成错误。

引语内部虽然可能很复杂，但从语法视角看，引语是引述语的宾语，这个宾语对于引述语是一个无需打开的硬核，因为引语内部的成分不会直接同引述语的成分发生句法关系或话头话体关系。一般来说，带有这一硬核的引述语构成一个小句复合体，硬核（即引语）内部则是一个或多个小句复合体。

根据这一观察，我们的标注方法是，

（1）把引述语和引语切分开，当做不同的小句复合体；

（2）在引述语中引语应当占据的宾语位置上加入一个带有索引的标识符，代表被共享的引语；

（3）引语由一个或多个小句复合体构成，其整体用上述标识符来索引。

如此标注的引述语和引语构成一个以引述语为核心、共享引语的小句复合体的组合。

例 1’

我们把例 1 划分成 4 个小句复合体。

A Lorillard spokeswoman said Q035,

Q035：

“This is an old story. .

We 're talking about years ago before anyone heard of asbestos having any questionable properties.

There is no asbestos in our products now .”

其中，Q035 是引语标识符，Q 表示引语，035 表示该引语在整个语篇的全部引语中的索引号。

A Lorillard spokeswoman said Q035, 是引述语，它是一个小句复合体。

Q035 是 said 的宾语，在引述语的小句复合体中是一个占位的硬核，不加分析。它本身的内容是 3 个小句复合体。

借助于对 Q035 的成分共享，这 4 个小句复

合体组合成一个整体。

这样的划分采用引语共享的理念，每一个小句复合体在句法和功能上都是完整的，引述语和引语的关系有清楚的标识，引述语不因引语的加入而功能过分复杂。这样的表示方法比起宾州树库和功能语法的划分，显然更合理了。

6 英汉小句对齐标注

6.1 标注体系

从大量语料中看出，英语小句复合体的汉语译文可以看成它的部件译文的组合，中间可能插入一些表示结构关系的汉语词，可能带有涉及指代关系和逻辑关系标记成分的变化。部件是英语小句复合体中的小句或者话头和话体，不再涉及更底层的语法单位。英汉小句对齐标注，就是在汉语译文中表现这种组合关系，为机器翻译、翻译教学、语言比较等应用提供这种结构化的数据。Ge & Song 提出了这一思想的轮廓[8]，这里制定了具体的做法：

第 1 步，基于引语共享来标注英语引述语和引语的小句复合体的组合（本文第 5 节）；用换行缩进图式表现英语小句复合体中的话头共享关系（本文第 4 节）。图式中的每一行或者是英语的小句，或者是英语小句的话头或话体。

第 2 步，给出英语小句复合体换行缩进图式中每行独立翻译的结果，在换行缩进图式里标注英语中被共享的话头在汉语译文中的对应成分。关系代词和关系副词以其全部字母大写形式充作译文，并在从句的译文中按照汉语句法语义要求置于适当位置。

第 3 步，以第 2 步的结果组合出英语小句复合体的整体汉语译文，该译文也用换行缩进图式列出，每一行是一个汉语标点句，话体标点句缩进到它的话头的右边，行尾标注出该行相对于独立译文行中成分的组合关系。

组合关系涉及的操作有 4 种：接续（带有调序），嵌入连接词语，成分的指代转换，逻辑连接成分的改变。限于篇幅，下面只是举例说明第 2 步和第 3 步标注方法的主要思想，标注规范详见[9]。

6.2 标注实例

例 2”

第 1 步，见例 2 的换行缩进图式即图 1。

第 2 步，图 1 中各行独立翻译结果如下：

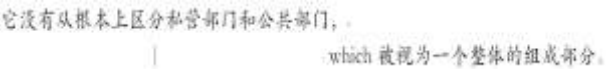

图 4 例 2 的换行缩进图式中各行独立翻译的译文

第 1 行右部的“私营部门和公共部门”是图 1 中英语第 1 行中 the private and the public sector 的汉语译文，英语的这个短语是图 1 中第 2 行的话头。图 4 中用换行缩进图式把这个话头关系标注出来。

第 3 步，把各行独立翻译的译文组合起来，成为整句译文（图 5）。

它没有从根本上区分私营部门和公共部门，//1
这两种部门被视为一个整体的组成部分。//sum(1.2)+*2

图 5 例 2 的整句译文

整句译文中各行内容是各行独立翻译的结果组合起来的。每行行尾双斜杠后面标注该行译文是如何组合的。

本例中，第 1 行最后双斜杠后边是 1，表示该行就是图 4 的第 1 行的内容。

图 4 第 1 行中“私营部门和公共部门”是第 2 行的话头，它左边的内容“它没有从根本上区分”与它分别记作 1.1 和 1.2。

在我们的标注体系中，数字 n 表示各行独立翻译结果的图中第 n 行的内容。第 n 行中可能会有被其他行共享的话头，这种共享话头的左右边界把这一行划分成几部分，这些部分自左至右分别用 n.1、n.2、n.3 等等表示。由于一行中可能有多个成分是被其它行共享的话头，所以划分后的部分可能不止 3 个。

sum(x)是一个函数，表示将它的自变量名词短语 x 做出概括，变成另一个同指的名词短语。这里自变量是 1.2，sum 就把“私营部门和公共部门”变成一个概括性的名词短语“这两种部门”。

第 2 行最后标注*2，指的是图 4 中第 2 行的内容，但要去掉关系代词 WHICH。

标注中的加号+表示把前后两部分内容连起来。sum(1.2)和*2 连起来就是“这两种部门被视为一个整体的组成部分。”

例 3”

图 3 中各行独立翻译结果如图 6。

一个人

WHO 享受过良好的教育机会，

应该在行为和语言上表现出来。

图 6 例 3 的换行缩进图式中各行独立翻译的译文

整句翻译结果是：

一个人享受过良好的教育机会，//1+*2

应该在行为和语言上表现出来。//3

图 7 例 3 的整句译文

其中第 1 行的组合关系 1+*2 表示把图 6 中的第 1 行与第 2 行（去掉关系代词 WHO）连起来，组合成整体译文的第 1 行。

例 4”

图 3 中各行独立翻译结果如图 8。

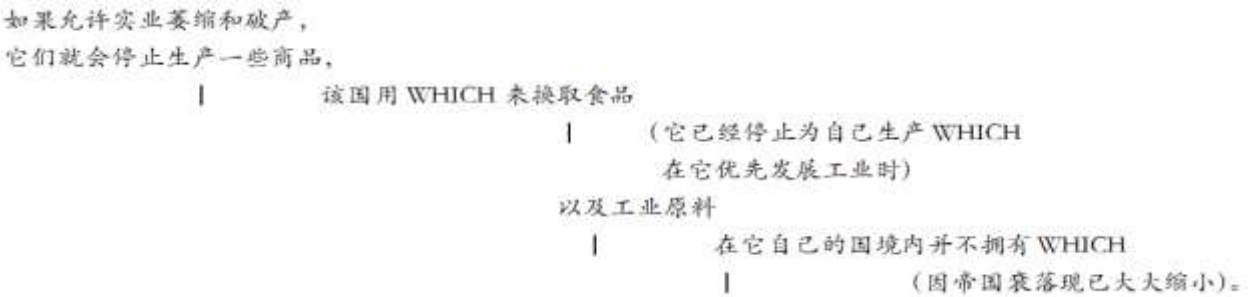

图 8 例 4 的换行缩进图式中各行独立翻译的译文

其中第 2 行被第 3 行的话头边界切分成 2 部分，2.1 是“它们就会停止生产”，2.2 是“一些商品”（第 3 行的话头）。

第 3 行被第 4 行的话头边界切分成 2 部分，3.1 是“该国用 WHICH 来换取”，3.2 是“食品”（第 4 行的话头）。

第 6 行被第 7 行的话头边界切分成 2 部分，6.1 是“以及”，6.2 是“工业原料”（第 7 行的话头）。

第 7 行被第 8 行的话头边界切分成 3 部分，7.1 是“在”，7.2 是“它自己的国境”（第 8 行的话头），7.3 是“内并不拥有 WHICH”。

组合成整句的译文见图 9。

如果允许实业萎缩和破产，//1

它们就会停止生产一些商品，//2

该国用这些商品来换取食品（它优先发展工业时它已经停止为自己生产），//rpw(3,det(2.2))+5+*4

以及在它自己的国境（因帝国衰落现已大大缩小）内并不拥有的工业原料。//6.1+7.1+7.2+8+*7.3+的+6.2

图 9 例 4 的整句译文

组合中第 1 行和第 2 行不变。

组合后第 3 行标注的 det(2.2)表示将图 8 中 2.2“一些商品”改成定指形式“这些商品”。rpw(x,y)表示用 y 取代 x 中的关系代词或关系副词，因此 rpw(3,det(2.2))就是用“这些商品”取代图 8 第 3 行中的 WHICH，于是得到“该国用这些商品来换取食品”，后面连上 5“它优先发展工业时”和*4“它已经停止为自己生产”，就得到组合后的第 3 行（括号另加）。

组合后的第 4 行标注 6.1+7.1+7.2+8+*7.3+的+6.2，表示把以下内容连在一起：“以及”（图 8 的 6.1）、“在”（图 8 的 7.1）、“它自己的国境”（图 8 的 7.2）、“（因帝国衰落现已大大缩小）”（图 8 的第 8 行）、“内并没有”（图 8 的*7.3）、“的”、

“工业原料”（图 8 的 6.2）。

本例也可以采用另一种译文，英汉对齐的整句译文及英汉小句对齐标注如图 10。

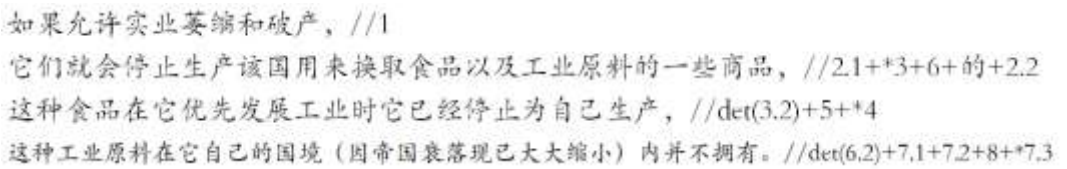

图 10 例 4 的另一种整句译文

7 讨论

现在常见的双语对齐语料库与本文论及的语料库相比，因目标不同，故原文选取、译文获取、标注方式都不一样。

通常的双语语料库目标是给数据驱动的机器翻译提供训练样本，本文语料库的直接目标是英汉小句复合体中小句语法结构对比，属于语言本体研究范畴，在此基础上再为机器翻译等应用目标服务。因此，本文的工作特别注重于理论基础的完善。本文以相当大的篇幅说明小句划分问题，便缘于此。

通常的双语语料库选取与应用领域相关的原文，或者通用性强的原文；本文语料库原文选取的是宾州树库用的华尔街日报语料，因为该语料语言比较规范，在领域和行文风格上具有多样性，从而具有语言研究价值，特别是有宾州树库句法树可以为翻译提供句法结构的参照。

通常的双语语料库译文是先于语料库就存在，是为目标语者阅读方便而翻译的；本文的语料库的原文语料没有公开发布的汉语译文，所以需要标注者自己翻译。翻译者主要为英语专业教师和研究生，并有宾州树库的句法树和多个机器翻译系统的翻译结果为参照，因此译文质量是比较可靠的。

通常的双语语料库的标注比较简单，一般是用统计方法自动求得双方大致的句对齐，以便低代价地得到海量的对齐文本，但其中会有噪音。本文的语料库的标注要体现英汉小句对照细节关系，因此由手工完成，力求完全准确。

需要特别说明的是，两种语言之间的翻译，译文并非只有一种可能。本文的语料库的翻译是人工完成的，人工翻译时做了选择，使得第 3 步的译文可以看成第 2 步译文的组合。这样的选择会省去翻译中的一些润色和变化，但可以反映出一个客观的语言事实：整体译文可以由部件译文组合而成，部件细分到源语言的话头话体为止。这是两种语言在小句复合体层面内的基本结构对应关系。至于具体的对应关系是怎样的，比如定语从句何时应译作先行语的左修饰成分，何时应另起一句；另起一句的情况下何时应共享话头，何时需要用某种形式复现话头，则是下一步的研究工作。本语料库已经为这种研究提供了结构化的标注样本。

这一语料库的标注展现了许许多多组合关系的实例，从而为归纳出组合的模式提供数据。当然，这种人工标注的语料规模远远达不到机器学习所需的训练语料的规模，尚不能为机器提供足够的训练样本，但是其中所表现的译文从局部到整体的组合关系，对于语言比较的理论研究和机器翻译的技术改进，都会有启发意义。特别是，目前机器翻译中深度学习的技术水平对于小句范围内上下文相关的翻译问题往往能处理得比较好，但小句复合体内小句间因话头共享而导致的上下文相关关系是相对远距离的，数据驱动的端到端的深度学习技术不易捕捉这种关系，该语料库则为此提供了信息。

8 结论

前人的语法理论对于语言单位的界定，特别是小句复合体和小句的界定，未提供面向话语的可操作的适当方法，难以支持自然语言处理的应用。本文为英汉小句对齐语料库建设做了扎实的理论准备，并在此基础上设计了标注体系和标注

步骤。

理论准备包括：

（1）界定了话语中的成分共享的概念，使用话头、话体和 NT 小句的概念进行英语小句复合体的分析。

（2）不受英语语法级阶关系的限制，以话头共享关系为基本线索界定小句。如此界定的小句具有功能的完整性和单一性，可以支持小句功能分析、小句间逻辑语义关系分析、不同语言小句复合体层面的小句对齐分析。

（3）分开引述语和引语，并在引述语中将引语表示为不加分析的句法单位，引述语通过宾语共享关系而表现完整的功能，从而将引语和引述语的分析纳入小句复合体分析的体系中。

操作设计包括：

（1）基于引语共享来标注英语引述语和引语的小句复合体的组合；用换行缩进图式表现英语小句复合体中的话头共享关系。

（2）将英语小句复合体到汉语译文的小句对齐标注工作分解为每行独立翻译和组合，设计了组合标注体系。

语料库的标注表明，在小句复合体层面上，英汉翻译涉及到的结构变换可以限制为英语的小句和话头、话体，无需涉及话头话体内部的结构。

这些工作对于语言本体研究和英汉语言对比、英汉机器翻译提供了结构化的标注样本。

参考文献

[1] Marcus M P, Marcinkiewicz M, Santorini B. Building a large annotated corpus of English: the Penn Treebank [J]. Computational Linguistics, 1993, 19(2): 313-330.

[2] Halliday, M.A.K. An Introduction to Functional Grammar (3rd Edition) [M]. London: Hodder Arnold, 2004.

[3] 王力. 王力全集 8：中国语法理论[M]. 北京：中华书局股份有限公司，2013.

[4] Kashioka H, Maruyama T, Tanaka H. Building a Parallel Corpus for Monologues with Clause Alignment [C] //Proceedings of the Ninth Machine Translation Summit, 2008.

[5] Koeva S, Rizov B, Stoyanova I, et al. Application of Clause Alignment for Statistical Machine Translation [C]// Proceedings of SSST-6, Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation. Jeju, Republic of Korea, 2012: 102–110.

[6] 王全蕊，李艳翠. 基于医学领域的汉英子句对齐语料库检索系统的设计与实现[J]. 河南科技学院学报(自然科学版)，2016，44（6）：57-62.

[7] 宋柔. 汉语小句复合体和话头结构[C]//揭春雨,刘美君. 实证和语料库语言学前沿. 北京:中国社会科学出版社，2018.

[8] Ge S & Song R. The Naming Sharing Structure and its Cognitive Meaning in Chinese and English [C] //Proceedings of SedMT 2016 (Workshop of NAACL 2016), Stroudsburg, PA: ACL, 2016: 13-21.

[9] 宋柔，葛诗利等. 英汉小句对齐语料库标注规范 V2.0 [R]. 广州:广东外语外贸大学外语研究与语言服务协同创新中心技术报告，2019

葛诗利（1969—），博士，教授，主要研究领域为计算语言学。
E-mail：geshili@gdufs.edu.cn

宋柔（1946—），通信作者，硕士，教授，主要研究领域为计算语言学。
E-mail：songrou@126.com